APPENDIX C

Mean

The sample mean is the most common measure of the central tendency in data. The sample mean is exactly the average value of the data in the sample. The implementation is as follows:

```
mean (N x 1 vector X) → y (scalar)
    1. Read in X
    2. X * = rm.missing(X) (removes records w/ missing values)
    3. N* = rows(X*)
    4. Call is.numeric(X*)
        a. If result is false then return error ‘Data must be numeric”
    5. Compute y using the following formula:
```

$$y = \sum_{i=1}^{N^*} X_i^* / N^*$$

```
    6. Return y
```

Max, Min, Median, Quartile and Percentile values characterize the sample distribution of data. For example, the α% of a data vector X is defined as the lowest sample value X such that at least α% of the sample values are less than X. The most commonly computed percentiles are the median (α =50) and quartiles (α =25, α =50 α =75). The interval between the 25th percentile and the 75th percentile is known as the interquartile range.

```
max.min (N x 1 vector X) -→ Y (2 X 1 vector containing min, max as elements)
    1. Read in X
    2. Remove missing and proceed (X now assumed non-missing )
    3. Call is.numeric
        a. If false then return error ‘data must be numeric’
    4. Set Y[1] =kth.smallest(1)
    5. Set Y[2] = kth.smallest(N)
    6. Return Y
median (N x 1 vector X) -→ y (scalar)
    1. Read in X
    2. Remove missing and proceed (X now assumed non-missing )
    3. Call is.numeric (see ads-other.doc)
        a. If result is false then return error ‘Data must be numeric”
    4. Compute k as the following:
        a. If N is even k = N/2
        b. Otherwise k = (N+1)/2
    5. Call kth.smallest(k)
    6. Return y = kth.smallest(k)
```

If N is even, statistics texts often report median as the average of the two ‘middle’ values. In one embodiment, the invention selects the *N/2-th* value. The reason is that with vary large data sets finding the computational time to find both values is often times not worth the effort.

percentile(*N x 1* vector X, *P X 1* vector Z containing the percentile values which must be between 0 and 1) -→ Y (*P X 1 vector containing percentiles as elements*)

Temporary Variables: Foo

```
1. Read in X
2. Remove missing and proceed (X now assumed non-missing )
3. Call is.numeric
   a. If false then return error 'data must be numeric'
4. Call is.percentage
   a. If false then return error 'percentile must be between 0 and 1'
5. For I = 1, …, P:
   a. Foo = floor(Z[I]*N)
         i. If Foo > 0 then Y[i] = kth.smallest(Foo)
        ii. Else Y[i] = kth.smallest(1)
6. Return Y
```

quartile(*N x 1* vector X) -→ Y (*3 X 1 vector containing quartiles as elements*)

Note: relies on percentile function (see above)

```
1. P = [0.25, 0.5, 0.75]
2. Y = percentile(X,P)
3. Return Y
```

Mode

The sample mode is another measure of central tendency. The sample mode of a discrete random variable is that value (or those values if it is not unique) which occurs (occur) most often. Without additional assumptions regarding the probability law, sample modes for continuous variables cannot be computed.

mode : (*N x 1* categorical vector X) -→ y (*scalar*)

```
1. Read in X
2. Remove missing and proceed (X now assumed non-missing )
3. Call is.numeric
   a. If result is false then return error 'Data must be numeric"
4. Call is.categorical
   a. If result is false then return error 'Data must be categorical'
5. Call array to hold list of unique objects, count for each object, and a scalar `MaxCount'
   variable to keep the current max count number in the array
6. Step through data and do the following:
   a. Check to see if object matches any object on current token list
         i. If yes
              1. Increment counter for that object by 1
              2. Check against MaxCount and increm. MaxCount if necessary
        ii. Otherwise,
              1. Create new list item and set count for this item to 1
              2. Check against MaxCount and increm. MaxCount if necessary
7. Check counts against MaxCount and return those items that match MaxCount (this will
   be at least one item but may be more than one ('bimodal', 'trimodal' sample
   distribution).
```

Sample Variance, Standard Deviation

The sample variance measures the dispersion about the mean in a sample of data. Computation of the sample variance relies on the sample mean, hence the sample mean function (see above) must be called and the result is referenced as μ_X in the following formula:

```
variance: (N x 1 vector X) -→ y (scalar)
  1. Read in X
  2. Remove missing and proceed (X now assumed non-missing )
  3. Call is.numeric (see ads-other.doc)
     a. If result is false then return error 'Data must be numeric"
  4. Call mean(X) and save result as $\mu_X$
     a. If mean(X) results in error then variance(X) returns error as well
  5. Compute y using the following formula:

        $\hat{\sigma}^2 = (1/(N-1))\sum_{i=11}^{N}(X_i - \mu_X)^2$

  6. Return y
stddev: (N x 1 vector X) -→ y (scalar)
  1. Read in X
  2. y = variance(X)
  3. y = sqrt(y)
  4. return y
```

Correlation

Correlation provides a measure of the linear association between two variables that is scale-independent (as opposed to covariance, which does depend on the units of measurement).

```
corr(N x 1 vector X, N x 1 vector Y) -→ z (scalar)
  1. Read in X, Y
  2. Remove missing and proceed (X, Y now assumed mutually non-missing – this means
     that all records where either x or y is missing are removed)
  3. Call is.numeric (see ads-other.doc)
     a. If result is false then return error 'Data must be numeric"
  4. Compute z using the following formula:
  $z = (1/N)\sum_{i=1}^{N}(X_i - \mu_X)(Y_i - \mu_Y)$
  5. Return z
```

Scenarios

The following example illustrates how these functions would be applied to a data vector X.

Let X = $(1, 3, 6, 11, 4, 8, 2, 9, 1, 10)^T$
mean X = 5.5
mode X = 1 (assuming here that these represent categories)
median X = 4
variance X = 13.05
stddev X = 3.6125